УДК 004.7:004.4

МЕТОД ОРГАНИЗАЦИИ СИСТЕМЫ ПОИСКОВОЙ РЕКЛАМЫ В СЕТИ ИНТЕРНЕТ

В.В. Силич

Томский государственный университет систем управления и радиоэлектроники
E-mail: acid@ms.tusur.ru

Предлагается метод организации системы поисковой рекламы, основанный на теории нечетких множеств, для выбора рекламных объявлений сайтов, релевантных поисковому запросу пользователя, при соблюдении ограничений рекламодателя.

Количество пользователей сети Интернет неизменно растёт год от года. При этом основная навигация в Интернете, как и все последние годы, осуществляется непосредственно с использованием поисковых машин. Но, несмотря на то, что на сегодняшний момент существует уже достаточное количество поисковых систем, они далеко не всегда способны удовлетворить информационные потребности пользователей. Как правило, это связано с несовершенством алгоритмов ранжирования, в соответствии с которыми поисковики определяют очередность, в которой пользователю будут выданы результаты поиска. И причиной этому является не неправильно построенные те или иные модели, или недостаточно отлаженные различные механизмы поисковиков. Основной причиной несовершенства поисковых систем является то, что они не понимают семантику самих запросов пользователей. Они могут при помощи сложнейшего алгоритма за секунды найти необходимые ключевые слова в миллионах документах, однако неспособны сопоставить эти данные с контекстом и смыслом самого запроса. Таким образом, в настоящее время дальнейшее совершенствование систем поиска в Интернете связано, прежде всего, с усилением семантической составляющей поиска, что позволило бы пользователям находить более релевантные документы, а не просто страницы, содержащие искомые ключевые слова [1].

Концепция Семантической паутины (Semantic Web) существует ещё с середины 90-х гг. XX в. Под данным термином подразумевается надстройка над существующей Всемирной паутиной (WWW), которая призвана сделать размещённую в сети информацию более понятной для компьютеров, а, следовательно, и для поисковых систем. В соответствии с данной концепцией, в сети Интернет каждый ресурс на человеческом языке должен быть снабжён описанием, понятным компьютеру. Однако данный подход всё ещё не имеет должного распространения и внедрения, так как по большей части предполагает отказ от существующих моделей строения Всемирной Сети и значительного изменения её структуры [2].

Таким образом, в настоящее время практически не существует эффективных механизмов, позволяющих программным образом найти в Сети искомую информацию, учитывая семантику самого запроса. В результате, разработчиками поисковых систем стали предприниматься попытки дополнить традиционные результаты поиска сайтами, которые были предварительно найдены по данному запросу другими пользователями системы и отмечены как особо релевантные. Для реализации подобной модели в результатах поиска пользователям предлагалось отметить те или иные документы как особо значимые и удовлетворившие информационную потребность пользователя по данному поисковому запросу. Такая схема позволяла дополнить машинные результаты поиска тем необходимым «смыслом», отсеяв большинство нерелевантных документов. Однако такая модель работы требовала, прежде всего, колоссальный объём накопленных пользовательских данных и предпочтений, ведь фактически, каждому поисковому запросу предварительно вручную должны были быть сопоставлены определенные результаты поиска. Помимо этого, результаты поиска, построенные на предпочтениях других пользователей, требовали и большого количества самих пользователей, заинтересованных в пополнении данной базы знаний поисковика, а такая заинтересованность по большей части отсутствовала. В итоге такая система «пользовательских» результатов поиска стала использоваться преимущественно лишь при узконаправленном тематическом поиске, где возможное количество результатов невелико, и ограничено [3].

Тем не менее, со временем появилась новая концепция для расширения результатов поиска. Наряду со стандартными результатами поиска, появились дополнительные результаты – «спонсорские ссылки» (сайты). Эти сайты также можно считать результатами поиска по определённому запросу пользователя, однако если традиционные результаты поиска выбираются поисковой системой на основе некоторых собственных критериев и алгоритмов, то «спонсорские» – на основе данных от самого владельца этого сайта. Т. е. сам владелец сайта или его доверенное лицо может внести в поисковую систему информацию о том, каким поисковым запросам будет релевантен его сайт. Таким образом, составляется ряд таких «рекламных» объявлений, представляющих собой ссылку на сайт и его краткое описание. Если в итоге пользователь перейдёт по такому объявлению – то рекламодатель произведёт некоторые финансовые отчисления за такой переход в пользу поисковой системы. Именно за счёт этого появляется заинтересованность владельцев сайтов в том, чтобы как можно точнее описать свой ресурс с точки зрения соответствия тем или иным поисковым запросам. В итоге поисковая

система накапливает базу данных сайтов и правил соответствия их поисковым запросам, которую постоянно пополняют рекламодатели, и которая будет полезна как рекламодателям, так и самим пользователям поисковика в качестве дополнительных результатов поиска. Система, организующая показ таких вот рекламных результатов поиска, называется «системой поисковой рекламы».

В общем виде современная система поисковой рекламы организована следующим образом. Имеется ряд пользователей поисковой системы, каждый из которых характеризуется набором поисковых запросов, которые он вводил в системе, историей сайтов, выбранных из результатов поиска и историей посещенных сайтов. Также существует ряд рекламодателей с множеством рекламных объявлений. Каждое рекламное объявление представляет собой ссылку на сайт и его краткое текстовое описание или графический баннер. Для каждого объявления задаются или наборы ключевых фраз – словосочетаний на естественном языке, или наборы сайтов, где данное объявление будет показано. Каждая ключевая фраза может состоять из одного или нескольких слов. Также среди ключевых фраз могут быть указаны, так называемые стоп-фразы, или по-другому анти-ключевые фразы. Соответственно для того, чтобы объявление было показано по тому или иному поисковому запросу, он должен содержать одну из ключевых фраз этого объявления и не содержать анти-ключевых фраз. Каждое из объявлений характеризуется определенной стоимостью, которую выбирает сам рекламодатель. В случае, если пользователь перейдёт на сайт по данному рекламному объявлению, эта стоимость будет списана со счета рекламодателя [4, 5].

При организации системы поисковой рекламы одной из основных задач, которые возникают ещё на стадии проектирования, является то, каким образом из множества спонсорских объявлений системы будет выбрано то подмножество объявлений, которое будет релевантно конкретному поисковому запросу пользователя. При этом выбранное множество объявлений (порядка 5–7 штук) должно также учитывать историю запросов и историю посещённых сайтов пользователем, а также ограничения самих объявлений. Эти ограничения выдвигаются самим рекламодателем, и могут быть связаны со стоимостью показа/перехода по объявлению, ограничениями на суммарный показ объявления за день и т. д. Рассмотрим один из способов решения данной задачи.

Искомое множество подходящих рекламных объявлений (релевантных запросу, а также дополнительным факторам и ограничениям задачи) можно определить как нечеткое. Следовательно, для решения поставленной задачи можно применить аппарат, используемый в теории нечетких множеств и нечеткой логики. Рассмотрим несколько упрощенный пример системы поисковой рекламы, основанной на текстовых объявлениях с размещением рекламы в результатах поиска на основе ключевых фраз.

Модель данной предметной области M включает множество рекламных объявлений системы $O=\{o_i\}$, текущий запрос пользователя fz, историю запросов пользователя Iz, историю выбранных пользователем сайтов Iv, историю посещенных сайтов Ip, а также информацию от пользователей с аналогичными предпочтениями A и ограничения рекламодателей R:

$$M = < O, fz, Iz, Iv, Ip, A, R > .$$

Каждое объявление системы можно представить следующим образом:

$$o_i = < h(o_i), t(o_i), l(o_i), v(o_i), \{fk_l(o_i)\}, \{fak_m(o_i)\}, s(o_i), p(o_i) >,$$

где $h(o_i)$ – заголовок объявления o_i, $t(o_i)$ – текст объявления o_i, $l(o_i)$ – ссылка (url-адрес) объявления o_i, $v(o_i)$ – видимый url-адрес объявления o_i, $fk_l(o_i)$ – ключевая фраза объявления o_i, $fak_m(o_i)$ – анти-ключевая фраза объявления o_i, $s(o_i)$ – стоимость объявления o_i, $p(o_i)$ – максимальное количество показов объявления o_i.

Каждый поисковый запрос пользователя fz представляет собой некоторую фразу, соответственно история запросов представляет собой множество запросов, вводимых пользователем ранее:

$$Iz = \{fz_k\}.$$

История выбранных пользователем сайтов из результатов поиска, как и история посещённых сайтов представляет собой множество ссылок (url-адресов) сайтов:

$$Iv = \{l_q\}, \quad Ip = \{l_s\}.$$

В терминах теории нечетких множеств представим совокупность подходящих объявлений системы (релевантных запросу и дополнительным факторам) как нечеткое множество $Op=\{<o,\mu_{Op}(o)>\}$, где o является элементом универсального множества или универсума O, а $\mu_{Op}(o)$ – функция принадлежности. Затем необходимо будет из этого множества выбрать некоторое количество наиболее подходящих объявлений On, которые и будут показаны в результатах поиска. Для этого определим On, как подмножество множества Op α-уровня [6]:

$$Op_\alpha = \{o \in O \mid \mu_{Op}(o) > \alpha\}, \quad \forall o \in O,$$

где $\alpha \in [0,1]$ и выбирается в соответствии с опытными данными системы.

Для представленных множеств будет справедлива следующая запись:

$$On \subseteq Op_\alpha \subset Op \subset O.$$

При этом мощность множества On не может быть больше максимального количества показываемых в системе объявлений N_{ob}, т. е.:

$$|On| \leq N_{ob},$$

где N_{ob} выбирается в соответствии с опытными данными системы и варьируется в целочисленном интервале [5,7], т. е. $N_{ob} \in [5,7]$.

Чтобы выбрать искомые 5–7 объявлений для показа в системе (подмножество *On*), будет достаточно выбрать из *Op* это количество объявлений с максимальными значениями функции принадлежности.

Таким образом, для каждого объявления нужно определить значение функции принадлежности множеству подходящих объявлений, т. е. степень того, насколько объявление соответствует информационным потребностям данного пользователя и ограничениям рекламодателя. Выделим основные признаки (факторы), по которым будет определяться принадлежность объявления множеству *Op*:

1. Соответствие поискового запроса ключевым фразам объявления.
2. Учёт истории запросов данного пользователя в системе.
3. Учёт сайтов, выбранных пользователем из результатов поиска.
4. Учёт сайтов, посещенных пользователем.
5. Учёт сайтов, посещаемых пользователями с аналогичными интересами/запросами.
6. Ценообразование объявления.
7. Частота показа объявления.

На рисунке приведена схема зависимости степени принадлежности объявления множеству подходящих объявлений от различных факторов.

По каждому признаку G_k, $k=\overline{1,7}$ определяется степень соответствия объявлений множеству *Op*, т. е. формируется своя функция $\mu_{G_k}(o)$ принадлежности объявлений множеству подходящих объявлений. Можно рассматривать признаки как критерии, по которым оценивается, насколько объявление является подходящим. Тогда интегральная оценка может определяться на основе методов свертывания критериев. Существуют различные методы свертывания. В случае «жесткой» постановки задачи («все или ничего») используются правила агрегации конъюнктивного или дизъюнктивного типа, которым соответствуют операции min или max, выполняемые над функциями принадлежности частных критериев. Если же стратегией интегральной оценки является компромисс, то используются различные операции осреднения [7].

В данном случае критерии (признаки) дополняют друг друга, причем важность их различна. Поэтому принадлежность объявлений множеству *Op* будем определять по формуле выпуклой комбинации нечетких множеств [7]:

$$\mu_{Op}(o) = \sum_{i=1}^{7} w_i \cdot \mu_{G_i}(o), \quad \sum_{i=1}^{7} w_i = 1,$$

где w_i – вес i-го признака. Веса признаков определяются с использованием метода «парных сравнений».

Функции принадлежности по различным признакам строятся различными способами. Рассмотрим формирование функции принадлежности по признаку «Соответствие поискового запроса ключевым фразам объявления», как самого важного, т. е. имеющего наибольший вес.

Рассматриваемый признак сам является составным, т. е. принадлежность объявления множеству *Op* по данному признаку при отсутствии в поисковом запросе стоп-фраз объявления, складывается из степеней соответствия поискового запроса *fz* каждой из ключевых фраз. В случае нахождения в поисковом запросе *fz* хотя бы одной из стоп-фраз множества $\{fak_l\}$, это объявление признаётся нерелевантным и исключается из дальнейшего рассмотрения. В противном случае следует анализ поискового запроса на соответствие ключевым фра-

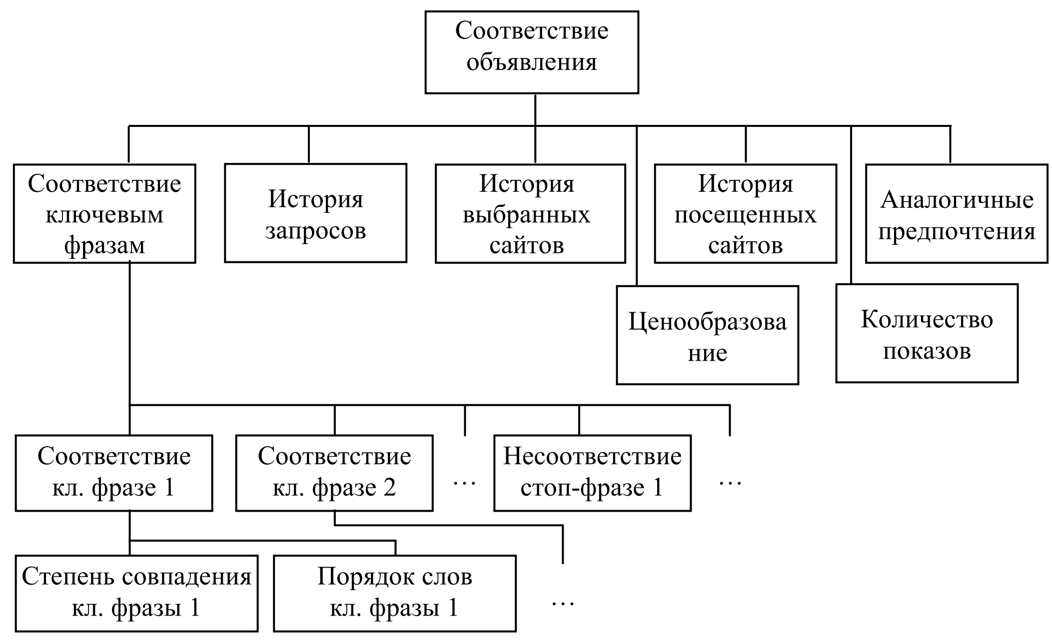


Рисунок. *Факторы, влияющие на принадлежность объявления множеству подходящих объявлений*

зам. Обозначим множество подходящих объявлений, определяемых по j-ой ключевой фразе, т.е. по соответствию fz и fk_j, через O_{fk_j}. Тогда множество подходящих объявлений O_{fk}, определяемое по всем ключевым фразам, зададим как объединение множеств O_{fk_j}:

$$O_{fk} = O_{fk_1} \cup O_{fk_2} \cup \ldots,$$

т. е. функция принадлежности определяется с помощью операции max:

$$\mu_{O_{fk}}(o) = \max_j \mu_{O_{fk_j}}(o).$$

Для определения функции $\mu_{O_{fk_j}}(o)$ необходимо для каждого объявления сравнить поисковый запрос fz и ключевую фразу $fk_j(o_i)$. Функцию $\mu_{O_{fk_j}}(o)$, отражающую степень соответствия запроса ключевой фразе, зададим аналитически следующим выражением:

$$\mu_{O_{fk_j}}(o) = \max(0, \frac{1}{n} \cdot \sum_{k=1}^{n} e_{kj} \cdot p_{kj}),$$

где n – количество слов запроса fz, e_{kj} – коэффициент, определяющий степень совпадения k-го слова ключевой фразы fk_j и запроса; p_{kj} – коэффициент, определяющий степень соответствия позиции k-го слова ключевой фразы fk_j по отношению к запросу. Коэффициенты e_{kj} и p_{kj} определяются по следующим формулам:

$$e_{kj} = \begin{cases} 1{,}0, & \text{если } k\text{-е слово фразы } fk_j \text{ есть в запросе полностью;} \\ 0{,}9, & \text{если } k\text{-е слово фразы } fk_j \text{ есть в запросе не полностью;} \\ -0{,}2, & \text{если } k\text{-го слова фразы } fk_j \text{ нет в запросе.} \end{cases}$$

$$e_{kj} = \begin{cases} 1{,}0, & \text{если позиция } k\text{-го слова } fk_j \text{ совпадает с позицией в запросе;} \\ 0{,}9, & \text{если позиция } k\text{-го слова } fk_j \text{ соответствует позиции в инверсном порядке;} \\ -0{,}2, & \text{если позиция } k\text{-го слова } fk_j \text{ не совпадает с позицией в запросе.} \end{cases}$$

В результате, применяя данные выражения для каждой из ключевых фраз объявления, можно рассчитать степень соответствия объявления текущему поисковому запросу по критерию «Соответствие поискового запроса ключевым фразам объявления». Аналогичным образом определяются функции принадлежности множества подходящих объявлений по другим признакам.

В целом, использование вышеописанного подхода при создании алгоритма поиска подходящего объявления в системе поисковой рекламы позволяет в рамках одной модели совмещать самые разные факторы, от которых зависит релевантность объявлений поисковым запросам и различного рода ограничениям. При этом каждому из факторов может быть определён свой весовой коэффициент, что позволяет легко корректировать степень влияния этих факторов на конечный результат.

СПИСОК ЛИТЕРАТУРЫ

1. Розенфельд Л., Морвиль П. Информационная архитектура в Интернете: 2-е изд. – М.: Изд-во «Символ-Плюс», 2005. – 544 с.
2. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2000. – 384 с.
3. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа. – М.: Изд-во «Диалектика», 2005. – 272 с.
4. Harold D. Google Advertising Tools. – Sebastopol: O'Reilly Media, 2006. – 366 p.
5. Long J. Google Hacking for Penetration Testers. – Rockland: Syngress Publishing, 2005. – 502 p.
6. Леоненков А.В. Нечеткое моделирование в среде MATLAB и fuzzyTECH. – СПб.: БХВ-Петербург, 2003. – 736 с.
7. Дюбуа Д., Прад А. Теория возможностей. Приложения к представлению знаний в информатике. – М.: Радио и связь, 1990. – 288 с.